

Hadronic Tau Lepton Tagging with the CMS detector using domain adaptation to mitigate discrepancies between simulation and data

Paola Mastrapasqua^{a,*} on behalf of the CMS Collaboration

*^aCentre for Cosmology, Particle Physics and Phenomenology (CP3),
UCLouvain, Louvain-la-Neuve, Belgium*

E-mail: paola.mastrapasqua@cern.ch

The DeepTau identification algorithm, based on a Deep Neural Network model, has been developed to reduce the fraction of jets, muons and electrons misidentified as hadronically decaying tau leptons by the hadron-plus-strip algorithm in CMS.

Its recently deployed version for the Run 3 of LHC, DeepTau v2.5, has brought several improvements to the existing algorithm, most importantly the inclusion of domain adaptation techniques specifically designed to reduce simulation-to-data discrepancies in the high-score region of the tagger. In this work, the main novelties of DeepTau v2.5 are briefly discussed and its improved performance are presented. The new model delivers a reduced jet fake rate by ~50% across the regions of interest and, thus, sets a new improved baseline for the tau identification task. Lastly, the calibration of the tagger is performed and proves that, as sought, the new version is less sensitive to simulation mismodelling.

*42nd International Conference on High Energy Physics (ICHEP2024)
18-24 July 2024
Prague, Czech Republic*

*Speaker

1. Introduction

Hadronically decaying taus (τ_h) are reconstructed with the hadrons-plus-strips (HPS) algorithm [1] in different decay modes (DMs) based on the number of charged hadrons and of clusters of electrons and photons forming a strip. The major challenge is the distinction between genuine τ_h , and quark and gluon jets, which are copiously produced in QCD multijet processes. For this purpose, machine learning techniques are exploited to build the DeepTau identification algorithm [2] that can efficiently identify genuine τ_h candidates and reduce the rate of jets, electrons and muons misidentified as τ_h . It exploits a combination of high level input variables and lower-level information from particles around the τ_h flight direction represented in a $\eta \times \phi$ grid format. The network outputs an estimation of the probabilities that a candidate is an electron, muon, τ_h , or quark/gluon jet (respectively y_e , y_μ , y_τ and y_{jet}). Finally, these probabilities are used to define the final three discriminators, one for each source of background: D_{jet} , D_e , D_μ , where $D_\alpha(\mathbf{y}) = \frac{y_\tau}{y_\tau + y_\alpha}$ with $\alpha \in \{e, \mu, jet\}$. More details on DeepTau v2.1 (first deployed version) can be found in Ref. [2].

2. Improved DeepTau algorithm (v2.5)

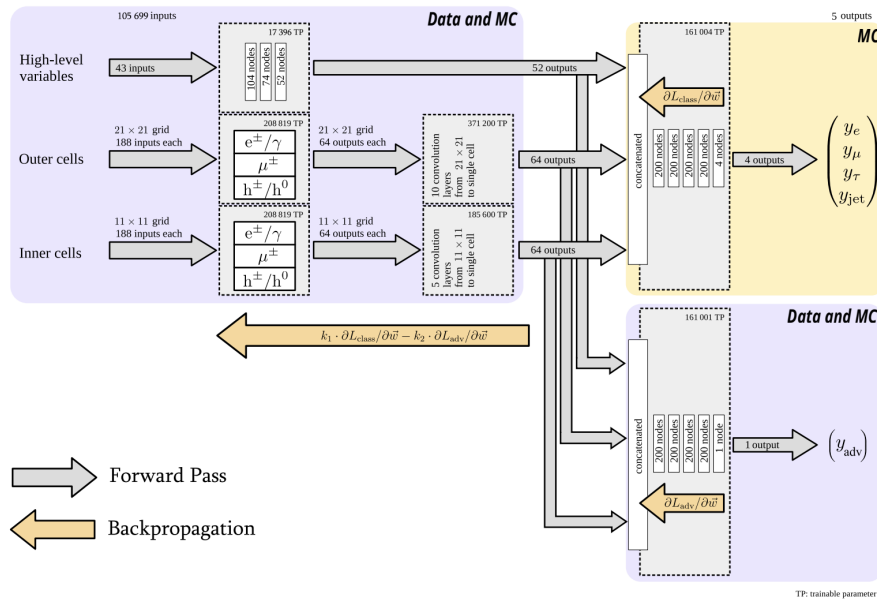


Figure 1: Diagram of the DeepTau v2.5 architecture, adapted from Ref. [2]. Each group of input variables (high-level features, inner cells and outer cells) are initially processed separately, then concatenated and passed through dense layers, before a final softmax layer which yields probability estimates that the candidate is an electron, muon, τ_h or jet. The subnetwork in charge of the classification is trained only on simulated data (yellow) while the adversarial subnetwork is trained with both simulated and real data (purple).

The DeepTau for the LHC Run 3 (v2.5) [3] is the improved iteration of its predecessor (v2.1). DeepTau v2.1 was trained exclusively on simulated Monte Carlo (MC) events, and while these generally provide a good representation of the pp collisions, some of the input features are not perfectly modelled. This led to increasing discrepancies between observed data and MC predictions

for high D_{jet} scores. This issue has been addressed in v2.5 by implementing domain adaptation using adversarial machine learning techniques. It involves simultaneously training two subnetworks with competing goals (Fig. 1). One is used for τ_h classification, and the other for domain discrimination between data and MC events. The latter is trained on a balanced mixture of simulated and real events with an average purity of $\sim 70\%$ obtained with $Z/\gamma^* \rightarrow \tau\tau$ selection. The overall goal is to maximise the classification performance while minimising data/MC discrimination. In addition, the new tagger is trained with updated detector conditions (2018), it integrates an improved *Shuffle and Merge* technique to balance events in p_T and η as well as a hyperparameter optimization and a revised input feature standardisation.

3. Performance and Calibration

The most challenging background source for τ_h identification are jets. As shown in Fig. 2 left, DeepTau v2.5 reduces the jet mis-id probability by $\sim 50\%$ with respect to v2.1. The calibration of the tagger is performed with a tag-and-probe technique on $Z/\gamma^* \rightarrow \tau_\mu\tau_h$ events. The resulting correction factors (shown in Fig. 2 right), closer to unity for v2.5, prove that the new architecture using domain adaptation techniques substantially reduces simulated-to-real data disagreement. More detailed results are shown in Ref. [4].

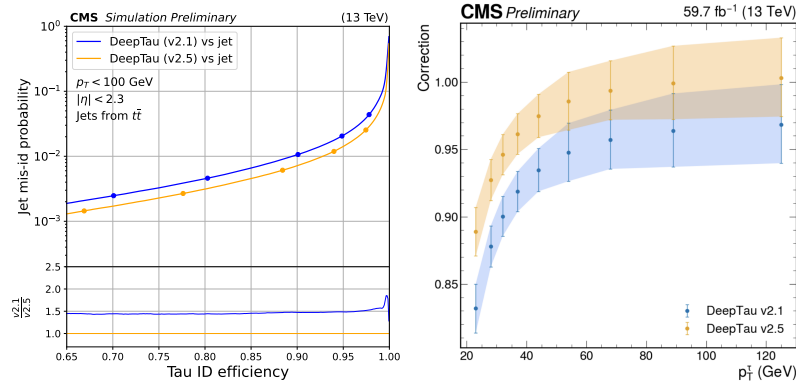


Figure 2: (Left) Jet misidentification probability versus tau identification efficiency evaluated on 2018 MC dataset. (Right) DeepTau v2.5 and v2.1 Tau ID correction factors evaluated on 2018 dataset.

References

- [1] CMS Collaboration, “Performance of reconstruction and identification of τ leptons decaying to hadrons and ν_τ in pp collisions at $\sqrt{s} = 13$ TeV”, *JINST* **13** (2018) P10005.
- [2] CMS Collaboration, “Identification of hadronic tau lepton decays using a deep neural network”, *JINST* **17** (2022) P07023.
- [3] L. Russell, “Identification of hadronic tau lepton decays with domain adaptation using adversarial machine learning at cms”, *CERN-THESIS-2022-126*.
- [4] CMS Collaboration, “Performance of the CNN-based tau identification algorithm with Domain Adaptation using Adversarial Machine Learning for Run 3”, *CMS-DP-2024-063*.